MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2010) doi: 10.1111/j.1755-0998.2010.02901.x

DNA BARCODING

The DNA Barcode Linker

MIHAI ALBU,*[1] HAMID NIKBAKHT,*[1] MEHRDAD HAJIBABAEI† and DONAL A. HICKEY*
*Department of Biology, Concordia University, 7141 Sherbrooke West, Montréal, QC H3G 1M8, Canada, †Biodiversity Institute of Ontario, Department of Integrative Biology, University of Guelph, Guelph, 50 Stone Road East, ON N1G 2W1 Canada

Abstract

DNA barcoding is based on the use of short DNA sequences to provide taxonomic tags for rapid, efficient identification of biological specimens. Currently, reference databases are being compiled. In the future, it will be important to facilitate access to these databases, especially for nonspecialist users. The method described here provides a rapid, web-based, user-friendly link between the DNA sequence from an unidentified biological specimen and various types of biological information, including the species name. Specifically, we use a customized, Google-type search algorithm to quickly match an unknown DNA sequence to a list of verified DNA barcodes in the reference database. In addition to retrieving the species name, our web tool also provides automatic links to a range of other information about that species. As the DNA barcode database becomes more populated, it will become increasingly important for the broader user community to be able to exploit it for the rapid identification of unknown specimens and to easily obtain relevant biological information about these species. The application presented here meets that need.

Keywords: DNA barcoding, bioinformatics, database searching

Received 5 January 2010; revision accepted 26 May 2010

Introduction

DNA barcoding is an innovative method for identifying biological specimens using a standard fragment of DNA sequence (Hebert *et al*. 2003; Hajibabaei *et al*. 2007; Burns *et al*. 2008). It can be used either to assign unknown specimens to known species or to provide preliminary evidence for the existence of new, cryptic species (i.e. Burns *et al*. 2008). DNA barcoding research can be divided into two phases. The first phase is the establishment of a list of reference DNA barcodes, based on DNA sequences from taxonomically verified voucher specimens. The second phase is the use of this reference list as a rapid method to assign new, unidentified specimens to a known species. Although the bulk of the research is still concerned with the first phase, it will become increasingly important to move the second phase in order for DNA barcoding to fulfil its promise. The method that we describe here is focused on the 'downstream' application of DNA barcoding by nontaxonomist end-users such as ecologists and conservation biologists.

We have exploited the power of the Google search engine (Singer & Hajibabaei 2009) to rapidly search through the database of DNA barcodes. To adapt this search engine to DNA sequences, we have broken the database sequences into a series of 'words', i.e. short subsequences of arbitrary length. Query sequences are also broken into words of equal length, and the Google Mini Search Appliance system is used to search the DNA barcode database. Once the species name has been retrieved based on the submitted sequence 'words' or characters, this name is used in turn to launch a series of other web-based searches to yield a large amount of information about that particular species. This allows the user to go from a short DNA sequence to a large body of biological information in a matter of seconds. In addition to the Google-based search, we provide the user with the option of also doing a search based on a customized version of the MEGABLAST program (Zhang *et al*. 2000).

The application is freely available at http://www.dnabarcodelinker.com.

Correspondence: Donal A. Hickey, Fax: 1-514-848-2881;
E-mail: donal.hickey@concordia.ca
[1]These authors contributed equally to this work.

User interface

The application includes a simple, self-explanatory graphical user interface that prompts the user to follow three simple steps.

1 First, the user pastes the query sequence into the search box and launches the application (see Fig. S1). The results page returns the species name (Fig. S1).
2 By clicking on the hyperlinked species name, the program returns a menu of further searches for that particular species (see Fig. S2)
3 By simply clicking on one of the icons shown in Fig. S2, the user can obtain a wide variety of specific information about that species (Fig. S2).

Database preparation

All sequences containing the BARCODE keyword were downloaded from NCBI and stored in our server and indexed by a Google Mini search appliance. These sequences were then broken into subsequence 'words' 120 bases long. In this way, a barcode of 600 bp can be broken into five 'words' of 120 bp each. For each DNA barcode sequence, this process was repeated using a sliding window approach, until words for 120 different frames were produced. For example, words in the first frame are bases 1–120, 121–240, 241–360, etc. Words in the second frame are bases 2–121, 122–241, 242–361, etc. This increases the size of the database by a factor of 120, but it means that the query sequences can be subdivided in a single arbitrary frame; this frame will match one of the 120 frames of the corresponding sequence in the database.

Query sequence submission

The query DNA sequence is submitted using a web-page interface. Either a single sequence or multiple sequences may be submitted. Multiple sequences are separated by a line containing the sequence label. This is essentially a simple FASTA format, although we do not assume the user is familiar with that format. Thus, we provide an online example of the sequence input requirements. The submission is then analysed by a PhP script, which identifies each sequence and breaks them into subsequence 'words' before comparing them to the database of DNA barcode sequences.

Output display

For each query sequence, there are three possible outcomes from the database search: (i) it matches the DNA barcode sequence for one of the species in the reference database; (ii) it matches DNA barcode sequences for more than one species in the database; and (iii) it does not match any of the sequences at the specified stringency, i.e. a perfect match over at least 120 bases. The initial output display indicates which of these categories corresponds to a given query sequence and provides a link for subsequent analysis.

In those cases where the query sequence matches the DNA barcode for a single species, the species name is shown as a hyperlink that leads to a second output page providing links to a wide range of biological information about that species. For example, this page provides automated links to Google Images, PubMed, and the Catalogue of Life. These links can easily be modified and updated as new resources, such as the Encyclopaedia of Life, become available. In the rare cases, when the query sequence matches DNA barcodes from more than one species, the names of all matching species will be displayed with links to detailed information about each species. A link to the Assigner program (Abdo & Golding 2007) is also provided, which allows the user to evaluate the likelihood that the query sequence comes from a particular species.

Finally, because the DNA barcode database is still far from complete, there will be cases where there are no DNA barcodes that closely match the query sequence. In these cases, the user will not be able to retrieve a species name. Instead, a link will be provided to the NCBI BLASTn server (Altschul *et al.* 1990) and a list of related sequences in the entire NCBI database (Barcodes sequences and other, nonbarcode sequences) will be returned. Because the BLASTn is less stringent, and the database extends beyond DNA barcode database, this allows the user to identify even distantly related DNA sequences. Thus, although a definite species assignment cannot be made in these cases, at least the user will be able to identify the major lineage to which the specimen belongs. Of course, as the DNA barcode reference database becomes more populated, this category of unassigned sequences will inevitably shrink.

Using MEGABLAST as an alternative search method

Because the main purpose of this web tool is to quickly link a specific DNA sequence to a species name and, through this species name, to link it to useful biological information, we chose to use a very simple search method. Other, more sophisticated, and more familiar search methods, such as BLAST (Altschul *et al.* 1990), are frequently used for comparing a query DNA sequence to a list of DNA sequences in a database. The power of the BLAST search lies in its ability to recover sequences with varying degrees of similarity to the query sequence and

to rank them by a similarity score. This method proved to be very powerful in earlier gene-finding studies. But because the COI sequence on which DNA barcodes are based is highly conserved, a BLAST search using any one DNA barcode sequence will recover thousands of other sequences from the database. Thus, the nonspecialist user is left to evaluate a list of BLAST scores. But it has been shown that the simple solution of choosing the top BLAST scores is not always reliable (Koski & Golding 2001), and one reason for this is that a longer sequence with an imperfect match may get a higher score than a shorter sequence with a perfect match to the query. This becomes a particular problem for DNA barcoding studies because the length of the sequences can vary depending on the primer design. Nevertheless, we have been able to customize the MEGABLAST variant of the BLAST program (Zhang *et al.* 2000) to perform essentially the same task as our simple Google-based search. MEGABLAST was originally developed to align almost-identical sequences that differed only by a couple of sequencing errors. Subsequently, it has become very useful for resequencing studies where there are occasional nucleotide differences because of both sequencing errors and naturally occurring DNA sequence polymorphisms. For our purposes, we also changed the default values for the MEGABLAST search, principally by restricting the search to sequences with the BARCODE keyword, increasing the word size from 28 to 128, and eliminating the masking of low-complexity sequences. The increase in word size makes the search more stringent and the taking into account of low-complexity sequences is necessary because regions of the mitochondrial DNA barcode sequences are frequently categorized by the MEGABLAST program as 'low complexity' because of the AT richness of mitochondrial DNA. We found that this modified MEGABLAST-based search gives essentially the same results as the simpler Google-based search, and we give the user the option of using both searches. If the MEGABLAST option is chosen, the results of both searches are shown side by side in the Results page (see Fig. S1).

Testing and benchmarking

Our search method is simple and fast because it does not require alignment of entire sequences, nor does it involve any model of phylogenetic inference. It simply picks sequence 'words' from a list in the database. Because these sequence words are quite long (120 bases), it is also very specific. We were concerned, however, that the method might be too selective and might not pick up a species name if the input sequence had minor differences from the database sequences (either because of real sequence polymorphisms within species, or because of occasional sequencing errors). To evaluate the system in cases where the query sequence differed from the database sequences, we did the following tests.

First, we used the Lepidoptera DNA barcode data set to benchmark our method. Lepidoptera have been the subject of several DNA barcoding studies (for example, Hajibabaei *et al.* 2006a) and they are represented in the DNA barcode database by more than 5000 DNA barcode sequences from 644 different species. The list of species names and the corresponding number of DNA barcode sequences is shown in Table S1. We took barcode sequences from each of the 644 species and used them to search the entire DNA barcode database. In all cases, the correct species was identified. We then asked if the correct species would still be identified if we discounted the barcode sequence that was used for the search. This test was performed for the 478 species which are represented in the database by more than a single DNA barcode sequence. With the exception of a single species, *Parnassius apollo*, other barcode sequences from the same species were identified in all cases. We investigated the exceptional case further, and we found that the two barcode sequences from this species show more than 2.5% sequence divergence, i.e. a greater divergence than is often seen between barcode sequences from different species. Thus, it is possible that we are dealing with a case of cryptic speciation in this instance.

In addition to returning the correct species, there were several cases where a second species (usually from the same genus) was also returned. These cases are as follows: *Adelpha melanthe* and *Adelpha phylaca*; *Aricia anteros* and *Aricia crassipuncta*; *Bungalotis astylos* and *Bungalotis midas*; *Callionima parce* and *Callionima denticulata*; *Cautethia yucatana* and *Cautethia spuria*; *Cobalus fidicula* and *Cobalus virbius*; *Leucanella memusae* and *Leucanella newmani*; *Myscelus assaricus* and *Myscelus perrissodora*; *Neoxeniades luda* and *Neoxeniades pluviasilva*; *Ornithoptera aesacus* and *Ornithoptera croesus*; *Parides eurimedes* and *Pyrrharctia isabella*; *Perigonia stulta* and *Perigonia lusca*; *Phyllonorycter heringiella* and *Phyllonorycter salictella*; *Polyommatus juno* and *Polyommatus eros*; *Saliana fusta* and *Saliana triangularis*; *Troides haliphron* and *Troides staudingeri*; *Xylophanes cthulhu* and *Xylophanes neoptolemus*; *Xylophanes lolita* and *Xylophanes loelia*. In all of these cases, however, the correct species is listed first in the results of our search, and the Assigner program (Abdo & Golding 2007) confirms that it is the more likely species.

There was one genus of Lepidoptera where both the Google-based method and the MEGABLAST-based method failed and that was within the genus Grammia. DNA barcode sequences from this genus routinely picked up more than two species from the same genus (see Fig. S3) and, in some of these cases, it was not possible to identify the correct species. On further investigation, we found that the reasons for the atypical results for Gram-

mia are twofold. First, different species share identical DNA barcode sequences and, second, there are species where the degree of sequence divergence between conspecific DNA barcodes is larger than the interspecific divergences. Indeed, this genus has been the subject of a publication (Schmidt & Sperling 2008) which concludes that DNA barcoding simply does not work as a species identification tool within the genus Grammia. In other words, the only case in which our search seemed to fail is the case where DNA barcoding itself is reported to fail.

In addition to benchmarking the method using existing DNA barcode sequences from Lepidoptera, we asked if the method would still perform well if presented with a novel sequence that differed from the existing database sequences by one or more nucleotides. For this test, we chose 10 DNA barcode sequences at random, from 10 different species (see Table 1). For each sequence, we introduced one, two, three random changes, etc., up to a maximum of 20 changes. Then, for each of these cases, we performed a search using the 'mutated' sequences. We repeated this process one hundred times, generating a different set of random changes each time. This resulted in a total of 2000 searches for each species, or 20 000 searches overall. For each set of 100 replicate searches, we asked what percentage of the time the correct species name was still returned. The complete results are shown in Fig. S4. Because our main concern was to determine approximately how many random changes would allow us to still recover the correct species name, this is summarized in Table 1. From the Table, we can see that the introduction of a few sequence variations did not prevent the system from accurately identifying the correct species. For example, as many as three or four random changes will still allow us to recover the correct species name 100% of the time. Thus, the system is tolerant of the normal level of intraspecific DNA barcode sequence variation, and it is not jeopardized by occasional minor sequencing errors. As a point of comparison, we used the variation within the DNA barcode sequences of *Cameraria ohridella*. This Lepidopteran species has been extensively sampled because of its importance as an invasive pest (Valade *et al.* 2009) and it is represented in the database by more than 500 DNA barcode sequences. Among all these sequences, the maximum divergence is three base pairs. Thus, our search seems well placed to deal with normal levels of intraspecific variation in DNA barcodes. Indeed, when we use any of the 500 DNA barcode sequences from *C. ohridella* to search the database, we still recover the correct species – even if we ignore the match to the query sequence itself.

Table 1 The number of random changes in a DNA barcode sequence that still allows the correct species name to be recovered from the DNA BARCODE database

Species name	GI number	100% species recovery	90% or greater species recovery
Venada cacao	84100697	4	5
Cocytius duponchel	288952332	4	7
Kloneus babayaga	284516226	4	6
Grammia obliterata	156618186	4	6
Gesta gesta	84098001	3	4
Falga sciras	288951074	4	5
Prosoparia floridana	284468514	4	5
Teinopalpus imperialis	145694515	3	5
Typhedanus undulatus	84100229	4	6
Enyo lugubris	288952376	4	6

Column 3 shows the maximum number of random changes that allowed the correct species name to be recovered in all 100 replicate trials. Column 4 shows the number of changes that allowed the correct name to be recovered at least 90 times of 100. For more details, see text and Fig. S4.

Validation

For all searches that return at least one species name, we provide the user with a link to the Assigner tool (Abdo & Golding 2007). This tool can be used to assess the statistical confidence of the species assignment. It is particularly useful in those cases where more than a single species name is returned by the search.

Discussion

Our search criteria are deliberately very stringent. To be scored as a 'hit', the query sequence must have a perfect match over 120 bases with a sequence in BARCODE database at NCBI. In practice, this means that we will retrieve only very closely related sequences from the database. In this way, we have minimized the number of false positives for species assignment. Interestingly, it has been shown that short DNA barcode sequences of about 120 base long – mini-barcodes – can reliably distinguish species and can be used in old specimens with degraded DNA, where a full-length DNA barcode cannot be sequenced (Hajibabaei *et al.* 2006a,b; Meusnier *et al.* 2008).

Setting highly stringent, search criteria might raise worries about the possibility of false-negative outcomes from the database searches. For example, if the stringency is too high, then the system might become too sensitive to occasional sequencing errors or to barcode query

sequences that are shorter than the standard DNA barcode. This is not, however, a problem with our method. This is because, instead of using the more usual alignment approaches over the whole sequence length, we use an alignment-free method based on subsequence 'words'. This means that the search is not severely compromised by either differences in query sequence length or by occasional sequencing errors that affect a minority of the sequence words.

Conclusions

The DNA Barcode Linker provides a user-friendly web interface that enables nonspecialists to exploit the potential of DNA barcoding for species identification. The program can be accessed at http://www.dnabarcodelinker.com.

Acknowledgements

This research was supported by grants from NSERC Canada and from Genome Canada (through the Ontario Genomics Institute) to DAH.

References

Abdo Z, Golding GB (2007) A step toward barcoding life: a model-based, decision-theoretic method to assign genes to pre-existing species groups. *Systematic Biology*, **56**, 44–56.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *Journal of Molecular Biology*, **215**, 403–410.

Burns JM, Janzen DH, Hajibabaei M, Hallwachs W, Hebert PD (2008) DNA barcodes and cryptic species of skipper butterflies in the genus *Perichares* in Area de Conservacion Guanacaste, Costa Rica. *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 6350–6355.

Hajibabaei M, Janzen DH, Burns JM, Hallwachs W, Hebert PD (2006a) DNA barcodes distinguish species of tropical Lepidoptera. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 968–971.

Hajibabaei M, Smith MA, Janzen DH, Rodriguez JJ, Whitfield JB, Hebert PDN (2006b) A minimalist barcode can identify a specimen whose DNA is degraded. *Molecular Ecology Notes*, **6**, 959–964.

Hajibabaei M, Singer GAC, Hebert PDN, Hickey DA (2007) DNA barcoding: how it complements taxonomy, molecular phylogenetics and population genetics. *Trends in Genetics*, **23**, 167–172.

Hebert PDN, Cywinska A, Ball SL, deWaard JR (2003) Biological identifications through DNA barcodes. *Proceedings of the Royal Society of London Biological Sciences*, **270**, 313–321.

Koski LB, Golding GB (2001) The closest BLAST hit is often not the nearest neighbor. *Journal of Molecular Evolution*, **52**, 540–542.

Meusnier I, Singer GAC, Landry JF, Hickey DA, Hebert PDN, Hajibabaei M (2008) A universal DNA mini-barcode for biodiversity analysis. *BMC Genomics*, **9**, 214.

Schmidt BC, Sperling FAH (2008) Widespread decoupling of mtDNA variation and species integrity in *Grammia* tiger moths (Lepidoptera: Noctuidae). *Systematic Entomology*, **33**, 613–634.

Singer GA, Hajibabaei M (2009) Googling DNA Sequences on the World Wide Web. *BMC Bioinformatics*, **10**, 14:S4.

Valade R, Kenis M, Hernandez-Lopez A *et al*. (2009) Mitochondrial and microsatellite DNA markers reveal a Balkan origin for the highly invasive horse-chestnut leaf miner *Cameraria ohridella* (Lepidoptera, Gracillariidae). *Molecular Ecology*, **18**, 3458–3470.

Zhang Z, Schwartz S, Wagner L, Miller W (2000) A greedy algorithm for aligning DNA sequences. *Journal of Computational Biology*, **7**, 203–214.

Supporting Information

Additional supporting information may be found in the online version of this article.

Fig. S1 Sample submission and result pages.

Fig. S2 Detailed information pages.

Fig. S3 Cross-matching of DNA barcodes between different species within the genus Grammia.

Fig. S4 Recovery of the correct species name using query sequences that have been modified to include varying numbers of random changes.

Table S1 Lepidopteran sequences in the DNA barcode database at NCBI

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting information supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.